Procesonderzoek en Padanalyse

Onderzoeksmethoden II

Inhoudsopgave

1 Voorspellen versus begrijpen

- Bij het *voorspellen van fenomenen* gaat het om het vinden van patronen ('correlaties') in de data. Op basis van data kunnen we voorspellen, beschrijven en correlaties ontdekken. Op basis van de data alleen kunnen we geen causale conclusies trekken.

- Om fenomenen te kunnen *verklaren* (begrijpen) is een causaal model nodig. In een causaal model wordt de kennis over oorzaken en gevolgen expliciet vastgelegd. Het causale model staat dus buiten de data en geeft weer wat de veronderstellingen zijn van de onderzoeker.

- Google kan bijvoorbeeld een griepgolf voorspellen op basis van zoektermen. Maar dat is geen causaliteit. Het is namelijk niet zo dat het zoekgedrag van mensen griep veroorzaakt. Dus als het zoekgedrag van mensen verandert als gevolg van deze berichtgeving, dan verandert het aantal griepgevallen niet.

- Het voorspellen van fenomenen met behulp van machine learning is de afgelopen jaren een groot succes gebleken. Grote hoeveelheden data en toegenomen computerkracht hebben geleid tot veel ontwikkelingen en toepassingen op het gebied van voorspellen. Voorbeelden zijn het real time voorspellen van beeldherkenning en het systeem van aanbevelingen bij bedrijven als Spotify, Bol, Netflix, ...

- De verwarring komt doordat het woord voorspellen in passieve, niet-causale zin gebruikt kan worden, bvb. als iemand ziek is en thuiszorg krijgt, voorspel ik een hoge kans dat deze persoon ouder dan 65 is. Maar het kan ook in actieve, causale zin gebruikt worden: bvb. ik voorspel dat als ik de prijs van dit product verlaag met 5%, dat ik er dan 10% meer van verkoop. Dit is een 'causale' voorspelling, oorzaak en gevolg.

- Het is cruciaal om onderscheid te maken tussen passief voorspellen en het voorspellen van oorzaak-gevolg effecten.

- Het feit dat 2 verschijnselen samen optreden betekent niet dat er een oorzakelijk verband is. Bij causaliteit gaat het om oorzaak en gevolg.

- Als een bepaald fenomeen de oorzaak is van een ander fenomeen, dan kunnen we ons ook counter factuals voorstellen. Dus wanneer het volgen van een opleiding leidt tot meer inkomen, dan kunnen we ook voor een gegeven persoon vragen wat zijn inkomen geweest zou zijn wanneer hij een andere opleiding gevolgd zou hebben.
- Helaas zien de data die we krijgen er vaak als volgt uit:

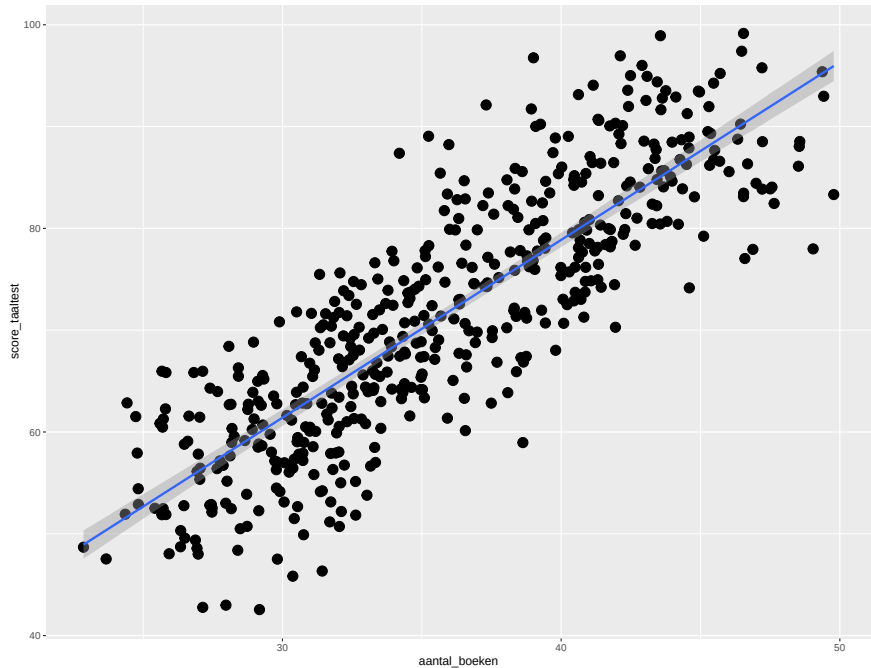
PP	X	Y_0	Y_1
1	0	67	NA
2	0	54	NA
3	1	NA	113
...	...	...	...

waarbij de uitkomst is weergegeven als Y en X staat voor het wel ($X = 1$) of niet ($X = 0$) doen van een interventie. Y_0 is de uitkomst wanneer er geen interventie is gedaan en Y_1 wanneer er wel een interventie is gedaan.

- Een causaal effect kan aangeduid worden als het verschil tussen de uitkomst wanneer een bepaalde interventie is toegepast, en de uitkomst wanneer de interventie niet zou worden toegepast.
- Het probleem is dat we meestal maar één van de potentiële uitkomsten kennen. De ontbrekende uitkomst is in de tabel weergegeven als 'NA' .
- Om echt causale conclusies te kunnen trekken, moet je goed nadenken over welke variabelen je wel en niet kunt gebruiken om Y gegeven de interventie X te kunnen voorspellen.

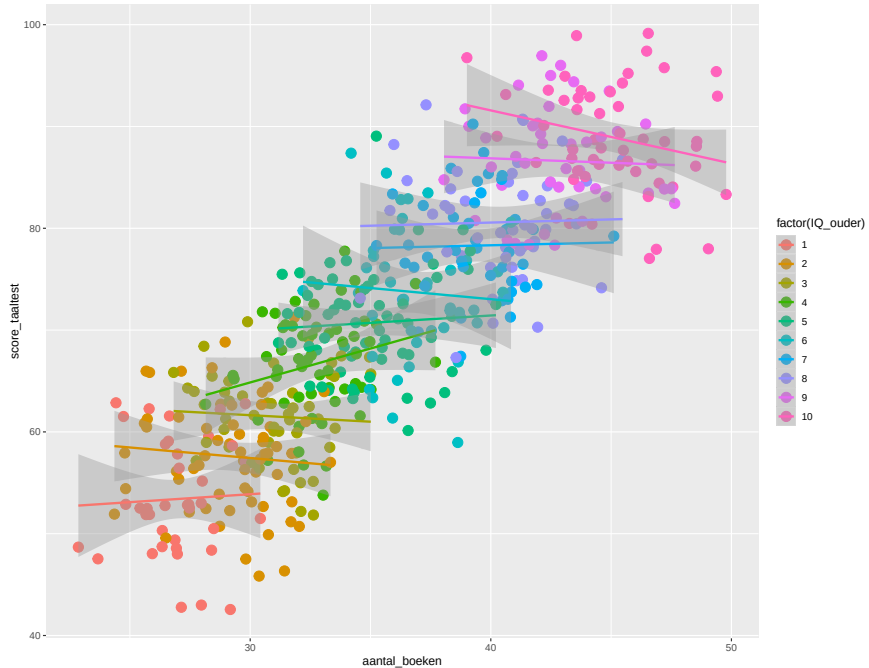
Taalvaardigheid bij kleuters

Om het probleem van het voorspellen op basis van correlaties te illustreren, bekijken we een fictieve dataset met gegevens over het aantal boeken dat ouders bezitten en de scores op taalvaardigheidstest bij 500 kleuters.



Er blijkt een correlatie te zijn tussen het aantal boeken dat ouders bezitten en de scores van hun kinderen.

- Als deze correlatie een causaal effect zou zijn, dan zou het helpen om boeken aan ouders van schoolgaande kinderen te geven.
- In dit geval is het aannemelijker dat er een andere variabele is, die zowel leidt tot meer boeken als tot hogere scores bij de kleuters: het IQ van de ouders.
- Wat gebeurt er als we rekening houden met ('controleren voor') het IQ van de ouders? Rekening houden met of controleren voor betekent in dit geval dat de variabele IQ van de ouders aan de analyse wordt toegevoegd. We kunnen bijvoorbeeld het IQ van de ouders in 10 even grote groepen indelen (op basis van percentielen bijvoorbeeld).

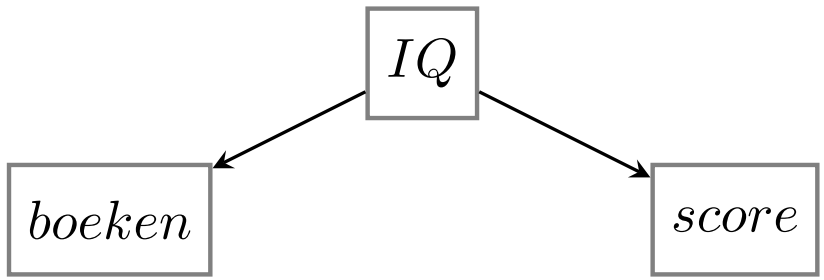


We zien dat binnen die groepen het aantal boeken de scores niet meer verklaren (vlakke lijn). Wanneer we informatie hebben over de scores en het IQ van hun ouders, voegt de data over het aantal boeken geen waarde meer toe aan onze analyse.

Het aantal boeken is geen oorzaak van de scores. Als we het aantal boeken zouden veranderen, dan verandert de score niet. Het IQ van de ouders is wel een oorzaak van hogere scores.

- Het is mogelijk om een causaal model m.b.v. vergelijkingen weer te geven. Dit wordt snel complex. Daarom worden causale modellen ook wel visueel weergegeven als *DAG's: Directed Acyclic Graphs*. Deze DAG's zijn een visuele weergave van een model dat alle oorzaak-gevolg relaties beschrijft.
- Een DAG bestaat uit 3 onderdelen:
  1. pijlen die de richting aangeven van causale relaties
  2. variabelen (die zowel geobserveerd als niet geobserveerd kunnen zijn)
  3. ontbrekende pijlen tussen variabelen

- We kunnen het hierboven gegeven voorbeeld over het aantal boeken dat ouders bezitten en de taalvaardigheidsscore van hun kleuters weergeven in een causaal model:

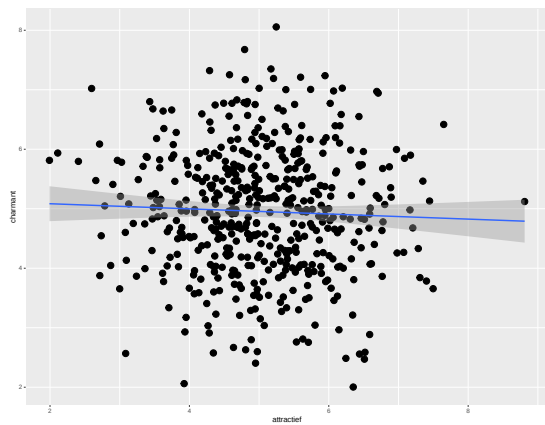


- Zoals uit het voorbeeld met de data bleek, maakt het uit of je *de confounder* (in het Nederlands: de gemeenschappelijke oorzaak) wel of niet opneemt in je analyse.

- Dus wanneer we uitgaan van de DAG, dan kunnen we alleen het causale effect identificeren wanneer we de confounder meenemen. Als we geen data zouden hebben over IQ ouders, dan kunnen we dus ook geen causaal effect schatten.

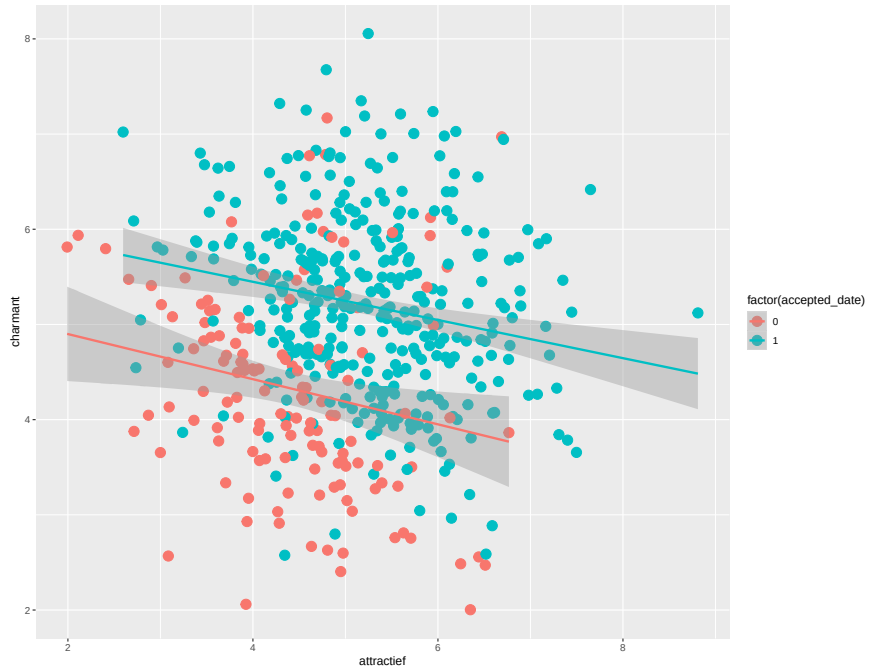
Potentiële partners: aantrekkelijk en charmant

- We willen de relatie identificeren tussen iemand aantrekkelijk vinden (‘attractief’) en iemand charmant vinden (‘charmant’) binnen een groep potentiële mannelijke partners van vrouwen. Daarbij gaan we ervan uit dat er zowel mannen zijn bij wie de vrouw een date heeft (‘accepted date=1’) en geen date heeft (‘accepted date=0’).
- We kunnen nu kijken wat er gebeurt wanneer we zowel daters als niet-daters meenemen in onze analyse.

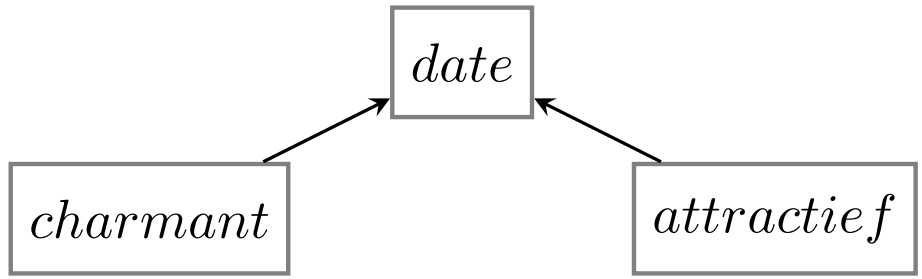


- We vinden geen correlatie tussen attractief en charmant.

- Maar wanneer we de dating status wel meenemen in de analyse, vinden we plotseling wel een verband!



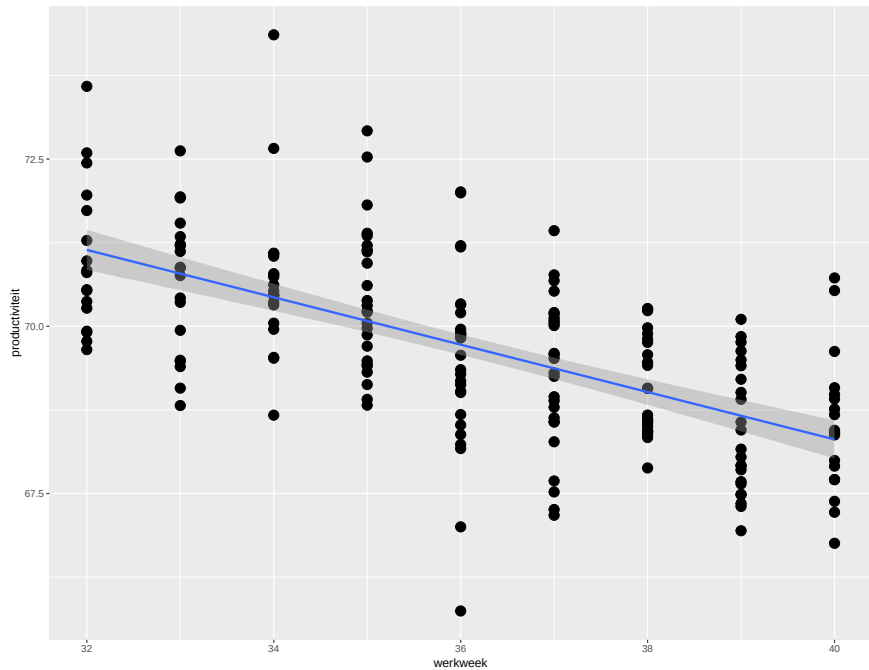
- Deze informatie kunnen we weergeven in de volgende DAG:



- Merk op dat er selectie plaatsvindt: We selecteren de daters (of juist de niet-daters), en na selectie ontstaat er een (negatieve) correlatie die er voorheen niet was.
- Stel dat een vrouw een date aanvaard heeft, maar de man niet aantrekkelijk vond, dan is het aannemelijk dat ze hem charmant vond (en vice versa). Door te conditioneren op date, krijgen we een negatieve correlatie, hoewel de 2 trekken in de volledige populatie mogelijks onafhankelijk zijn.
- We spreken van ‘collider bias’ of ‘selection bias’. We moeten colliders dus nooit meenemen in analyses. Wanneer we dat wel doen, dan trekken we ten onrechte de conclusie dat er een causaal verband bestaat.

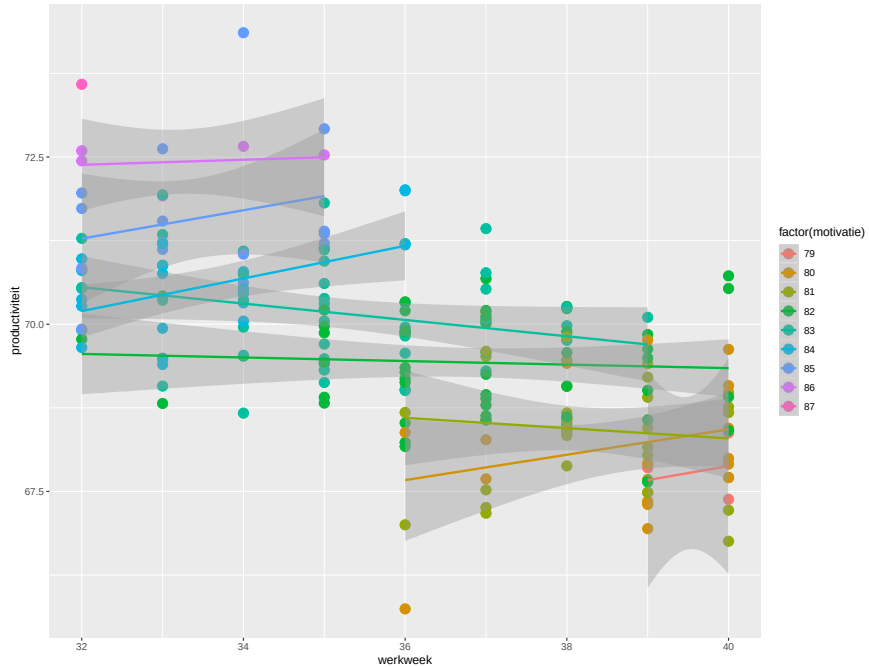
Leidt verkorting van de werkweek tot hogere productiviteit?

- Bij 200 werknemers met verschillende duur van de werkweek werd de productiviteit gemeten.



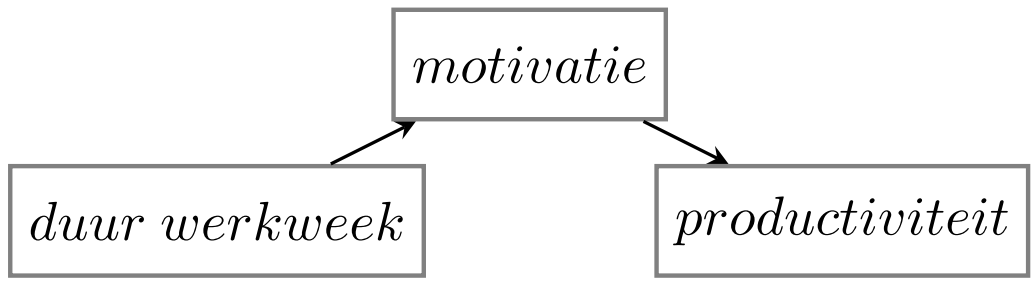
- Hoe korter de werkweek, hoe hoger de productiviteit?
- Welk mechanisme kan dit effect verklaren?

- Kan motivatie een verklaring zijn?



- Hoe korter de werkweek, hoe hoger de motivatie. Hoe hoger de motivatie, hoe hoger de productiviteit.
- Controleren voor motivatie, doet het effect van werkweek op productiviteit helemaal verdwijnen.

- Deze 'tussenliggende variabele' op het causaal pad van duur werkweek op productiviteit wordt de *mediator* genoemd.
- Indien de mediator een volledige verklaring biedt voor het onderliggend proces, kunnen we de DAG als volgt voorstellen



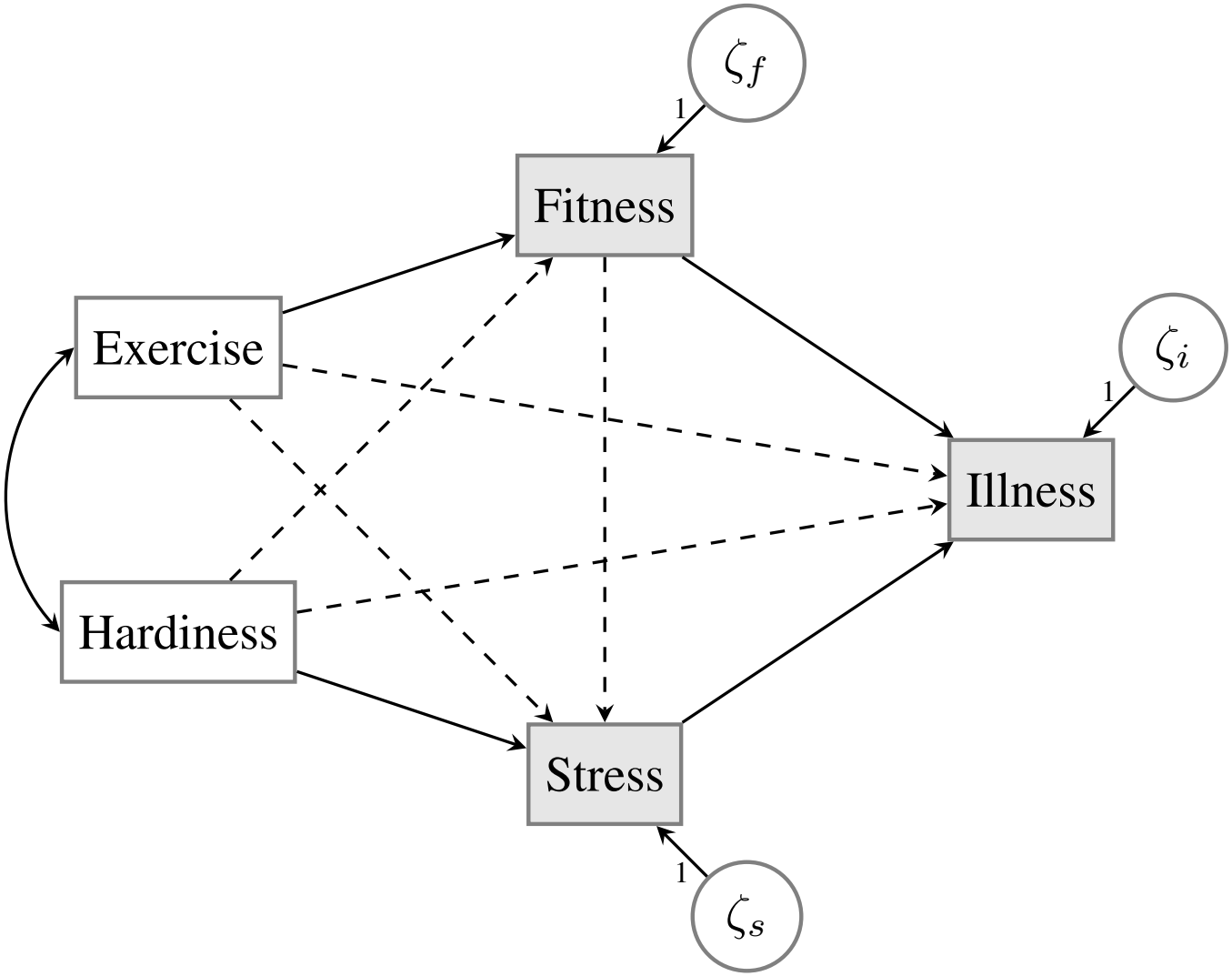
- Met de 3 bovenstaande voorbeelden hebben we laten zien dat het essentieel is om een causaal model te maken als je causale effecten wilt schatten. Soms is de verleiding groot om zoveel mogelijk variabelen mee te nemen in een model. Zoals net aangetoond kan dit tot verkeerde conclusies leiden!

2 Padanalyse

De impact van fysieke activiteit en veerkracht op ziekte

- Metingen voor exercise, hardiness, fitness, stress, illness in een steekproef van 373 university studenten
  - illness: mate van fysieke ziektesymptomen (laatste maand)
  - stress: mate van stressvolle 'life events' (laatste maand)
  - fitness: 'self-perceived physical fitness'
  - exercise: 'current exercise activity' (participatie in fysieke oefeningen)
  - hardiness: 'dispositional traits such as resiliency and willingness to look for opportunities in difficult situations'
- Direct of indirect effect van exercise/hardiness op illness?

Op basis van theorie stellen de onderzoekers onderstaand model voor:

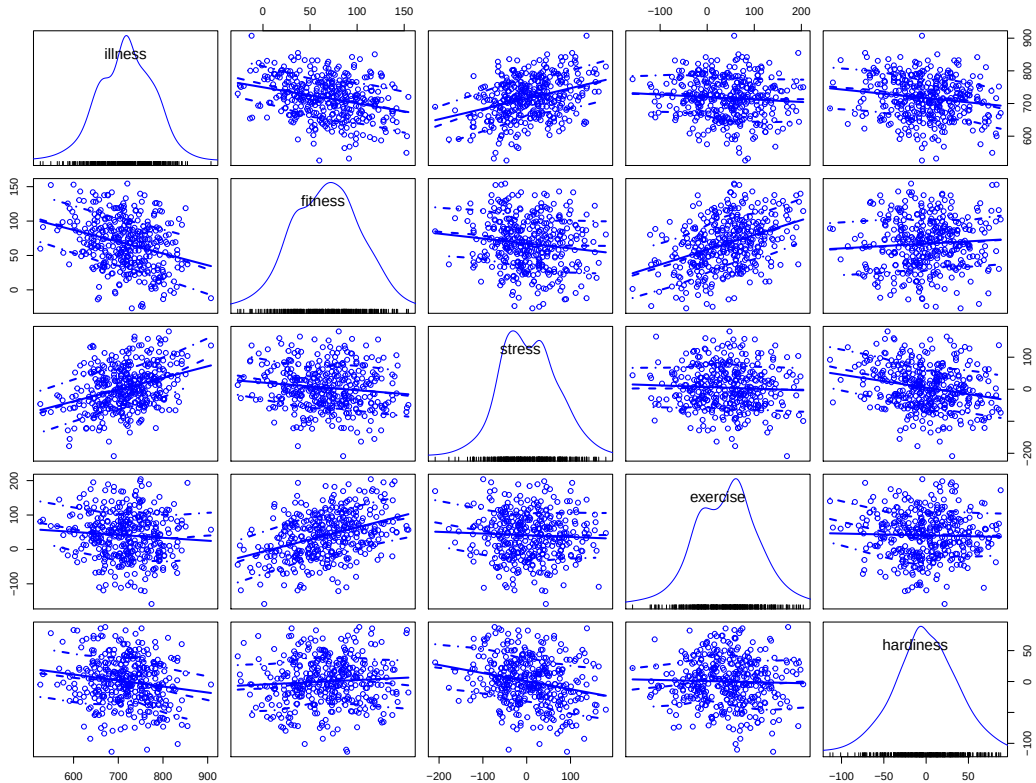


Padanalyse: analyse van structurele modellen waarbij alle variabelen geobserveerd (manifest) zijn.

- een *structureel model* representeert de hypotheses omtrent de patronen van directe effecten tussen deze variabelen
- elk theoretisch construct in het model (bvb. stress, depressie, welbevinden, openheid, ...) wordt 'gemeten' door 1 geobserveerde variabele, en correspondeert dus met 1 variabele in de dataset

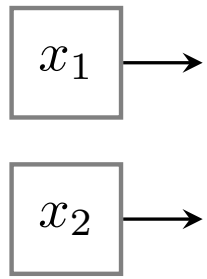
Padanalyse is onderdeel van SEM: 'Structural Equation Modeling' (structurele vergelijkingsmodellen)

Scatterplotmatrix van de data



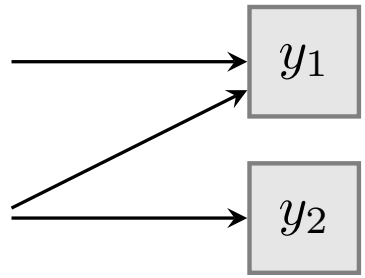
2.1 Elementen van een padanalyse model

1. geobserveerde *exogene* variabelen:



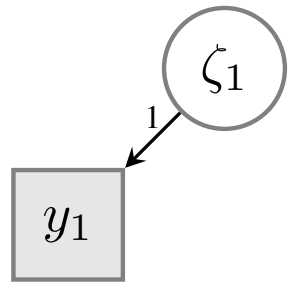
- de oorzaken van exogene variabelen worden niet verklaard binnen het model: bij een exogene variabele komt er nooit een enkele pijl toe
- exogene variabelen mogen covariëren (correleren), doch er wordt geen antwoord gegeven op de vraag waarom deze variabelen correleren (bvb. gemeenschappelijke oorzaak, de ene veroorzaakt de andere, . . .)

2. geobserveerde *endogene* variabelen:



- de (veronderstelde) oorzaken van endogene variabelen maken expliciet deel uit van het model
- padanalyse poogt een verklaring te bieden waarom endogene variabelen (al dan niet) correleren met elkaar
- bij elke endogene variabele komt er minstens één enkele pijl toe

3. niet-geobserveerde (latente) exogene variabelen (i.e. *disturbance* term)

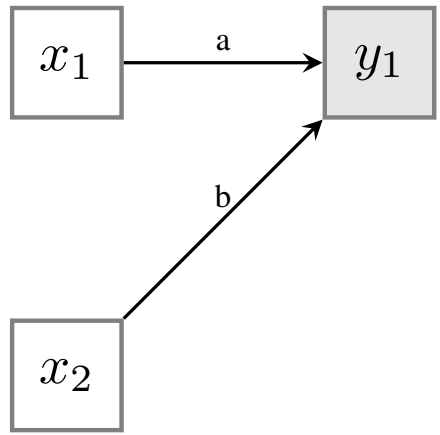


- elke endogene variabele heeft een *disturbance* term: deze disturbance term representeert alle weggelaten of niet-gekende oorzaken
- de (geschatte) variantie van deze disturbance term $\mathsf{Var}(\zeta_1)$ is een maat voor de proportie *niet-verklaarde* variantie van de endogene variabele y_1:

$$\mathsf{Var}(\zeta_1) = (1 - R^2)\mathsf{Var}(y_1)$$

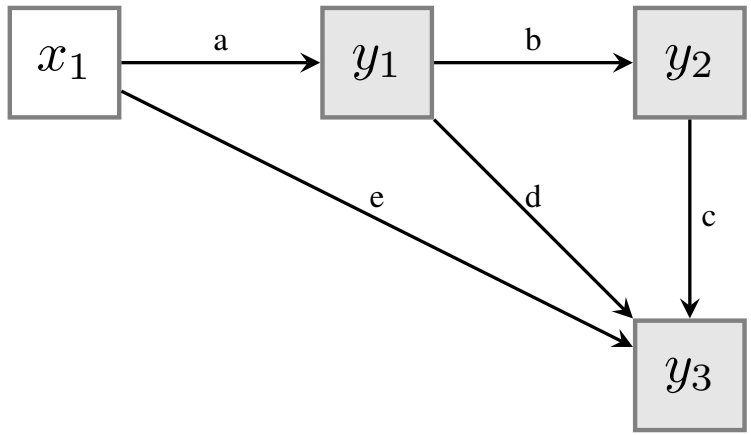
met R^2 de proportie verklaarde variantie van y_1 (= de determinatie-coëfficiënt op basis van de regressie van y_1 op alle variabelen die een effect hebben op y_1)

4. (verondersteld) directe effecten: enkele pijl



- de schatting van een *direct* effect is een *pad coëfficiënt*
- analoog met regressiecoëfficiënten
- het effect is lineair

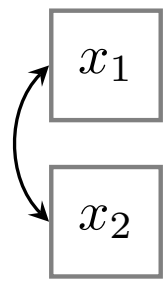
5. door de opeenvolging van directe effecten kunnen in een padanalyse op een natuurlijke wijze *indirecte* effecten worden gemodeleerd



- e representeert het *direct* effect van x_1 op y_3
- het *indirect* effect van x_1 op y_3:

$$(a \times b \times c) + (a \times d)$$

6. varianties van de exogene variabelen (zowel geobserveerde variabelen, als de latente disturbances)
    - worden doorgaans niet expliciet aangeduid op het paddiagram
7. covarianties tussen de exogene variabelen (dubbele pijl)
    (a) tussen twee geobserveerde variabelen:



    (b) tussen twee disturbance termen (standaard niet verondersteld)

2.2 Padanalyse a.d.h.v. lineaire regressie

Het effect van ‘exercise’ en ‘hardiness’ op ‘fitness’:

```
lm(formula = fitness ~ exercise + hardiness, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-86.148 -23.007  -0.201  22.954  87.889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.21746    2.05826  28.285  < 2e-16 ***
exercise     0.21718    0.02640   8.227 3.31e-15 ***
hardiness    0.07919    0.04620   1.714   0.0873 .
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 33.84 on 370 degrees of freedom
Multiple R-squared: 0.1588,     Adjusted R-squared: 0.1542
F-statistic: 34.92 on 2 and 370 DF,  p-value: 1.283e-14
```

- significant effect van exercise op fitness $(b = 0.21, s.e. = 0.03, t(370) = 8.23, p < .001)$
- 15.9% van de variabiliteit in fitness wordt verklaard door exercise en hardiness

Het effect van 'exercise', 'hardiness' en 'fitness' op 'stress':

```
lm(formula = stress ~ hardiness + exercise + fitness, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-196.18  -48.41   -1.70   43.24  163.55

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.68458    7.03082   2.658  0.00821 **
hardiness   -0.39284    0.08909  -4.409 1.36e-05 ***
exercise    -0.01434    0.05515  -0.260  0.79502
fitness     -0.19818    0.09986  -1.985  0.04794 *
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 65.01 on 369 degrees of freedom
Multiple R-squared: 0.06611,    Adjusted R-squared: 0.05852
F-statistic: 8.707 on 3 and 369 DF,  p-value: 1.355e-05
```

- significant effect van hardiness $(b = -0.39, s.e. = 0.09, t(369) = -4.41, p < .001)$ en fitness $(b = -0.20, s.e. = 0.10, t(369) = -1.99, p = .048)$ op stress
- 6.6% van de variabiliteit in stress wordt verklaard door exercise, hardiness en fitness

Het effect van 'exercise', 'hardiness', 'fitness' en 'stress' op 'illness'

```
lm(formula = illness ~ fitness + stress + exercise + hardiness,
    data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-182.404  -36.813    1.845   38.803  146.221

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 743.74194    6.19749 120.007  < 2e-16 ***
fitness      -0.44177    0.08766  -5.040 7.33e-07 ***
stress        0.27125    0.04545   5.967 5.67e-09 ***
exercise      0.03176    0.04816   0.659    0.510
hardiness    -0.12146    0.07981  -1.522    0.129
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 56.76 on 368 degrees of freedom
Multiple R-squared: 0.1835,     Adjusted R-squared: 0.1746
F-statistic: 20.67 on 4 and 368 DF,  p-value: 2.208e-15
```

- significant effect van fitness $(b = -0.44, s.e. = 0.09, t(368) = -5.04, p < .001)$ en stress $(b = 0.27, s.e. = 0.05, t(368) = 5.97, p < .001)$ op illness
- 18.4% van de variabiliteit in illness wordt verklaard door de 4 predictoren

2.3 Lineaire regressie versus covariantie-gebaseerde SEM

- ANOVA/lineaire regressie:
  minimalizeer het verschil tussen geobserveerde en voorspelde individuele uitkomst (kleinste kwadratenmethode)
  - enkel voor modellen met ongecorreleerde endogene variabelen
  - de analyse komt neer op een serie van multiple regressie analyses:
    * telkens één endogene variabele als afhankelijke variabele
    * alle variabelen die een direct effect hebben op deze endogene variabele beschouwen we als predictoren
    * de bekomen regressiecoëfficiënten zijn meteen de padcoëfficiënten
    * de schatting van de error-variantie (σ^2_ϵ) is meteen een schatting van de disturbance varianties voor deze endogene variabele ($\mathsf{Var}(\zeta)$)

- Covariantie-gebaseerde structurele vergelijkingsmodellen:
  minimalizeer het verschil tussen geobserveerde en voorspelde covariantie tussen variabelen
  - de meest courante schattingsmethode: 'maximum likelihood estimation'
    - vaak exact dezelfde parameterschattingen als met multiple regressie methode
    - assumptie: endogene variabelen multivariaat normaal verdeeld
    - alle parameters worden terzelfdertijd geschat
    - de schatting is iteratief: start met initiële waarden die na elke iteratie worden geupdatet
    - vertrekt van de variantie-covariantie matrix (en niet de correlatie matrix)

3 Het schatten van de parameters in padanalyse

3.1 Vrije parameters, datapunten en vrijheidsgraden

De vrije parameters in een padanalyse model

Standaard worden de exogene variabelen als vast beschouwd (fixed.x=TRUE), en zijn dit de parameters die geschat moeten worden:

1. de padcoëfficiënten van de directe effecten
2. de varianties van de disturbance termen
3. de covarianties tussen de disturbance termen (zeldzaam)

Alternatief worden ook de (co)varianties van de exogene variabelen (fixed.x=FALSE) als extra vrije parameters beschouwd.

Het aantal datapunten in een padanalyse model

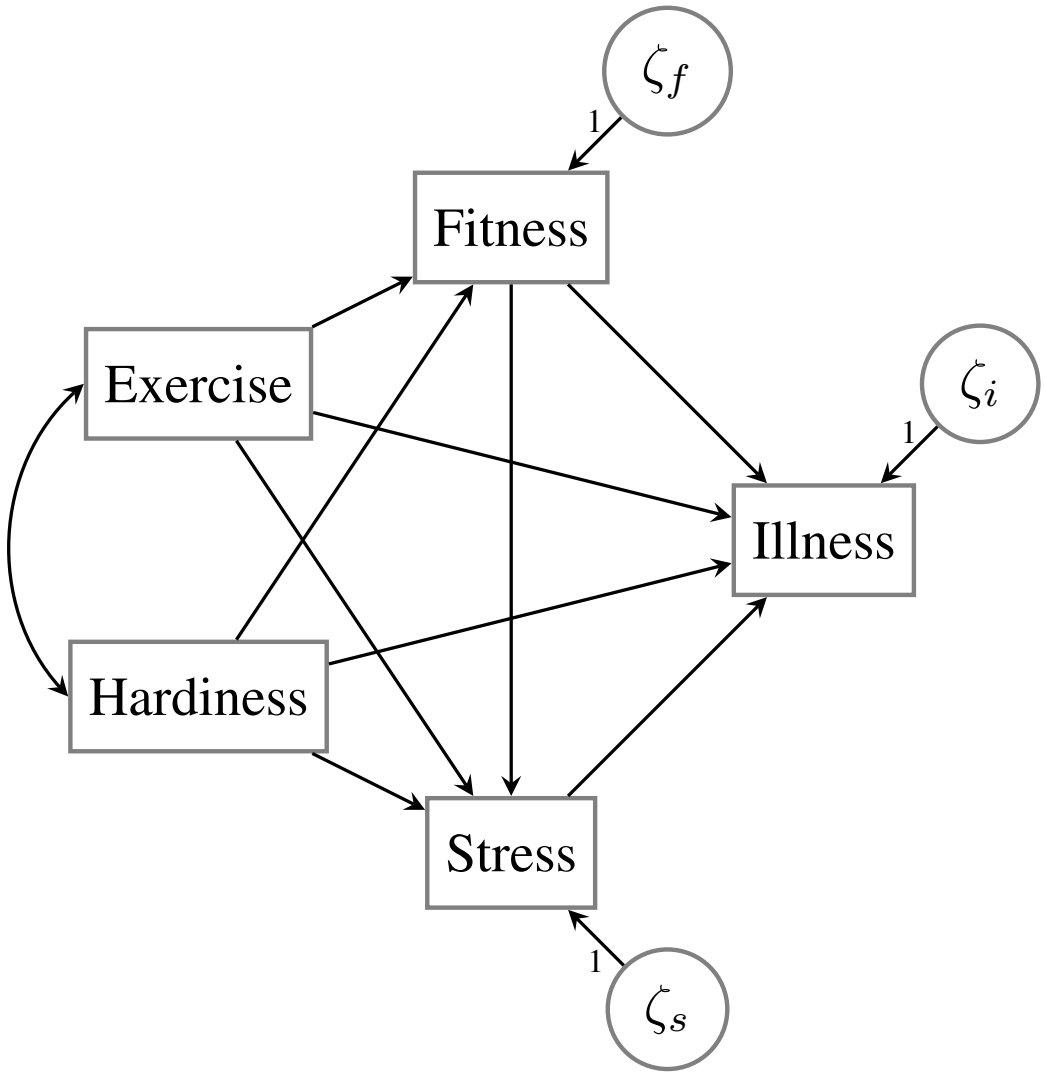
In een padanalyse berekent men het aantal datapunten op basis van het aantal (niet-redundante) elementen in de covariantie-matrix van de variabelen.

- indien er p geobserveerde variabelen in het model worden betrokken zijn er p varianties, en $p(p-1)/2$ covarianties, of samen $p(p+1)/2$ elementen
- in de 'fixed.x=TRUE' benadering worden de (co)varianties van de exogene variabelen niet meegerekend als datapunten; indien er q exogene variabelen zijn, zijn er slechts $p(p+1)/2 - q(q+1)/2$ datapunten
- het aantal datapunten blijft gelijk indien er meer observaties (subjecten) worden toegevoegd aan de dataset
- het aantal parameters van een model kan niet groter zijn dan het aantal datapunten waarop de analyse wordt uitgevoerd: het model is *niet-geïdentificeerd*

Vrijheidsgraden

- het verschil tussen het aantal datapunten en het aantal vrije parameters in het model noemt men de *vrijheidsgraden* [Engels: degrees of freedom (df)]
- indien het aantal parameters exact gelijk is aan het aantal datapunten, is het model *gesatureerd* (net geïdentificeerd), en zal de *fit* van het model perfect zijn; niettemin blijft de interpretatie van de parameters zinvol

Voorbeeld: aantal parameters en aantal datapunten



aantal parameters en aantal datapunten (fixed.x=TRUE)

- aantal parameters: 12
  - 9 padcoëfficiënten
  - 3 disturbances (residuele varianties)
- aantal datapunten: $p = 5$ en $q = 2$, dus $p(p+1)/2 - q(q+1)/2 = 12$
- het model is volledig gesatureerd (df=0)

3.2 Parameterschatting

- we zoeken die waarden voor $\boldsymbol{\theta}$ zodat het verschil tussen wat men in de data observeert ($\mathbf{S}$) en wat het model impliceert $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ zo klein mogelijk is
- verschillende manieren om dit 'verschil' uit te drukken leidt tot verschillende discrepantiematen
- de meest gebruikte discrepantiemaat is gebaseerd op maximum likelihood:

$$F_{ML}(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}| + \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - p$$

  waarbij p het aantal geobserveerde variabelen (trace is de som van de elementen op de diagonaal van de matrix)
- in de praktijk wordt tijdens het schatten $\boldsymbol{\Sigma}$ vervangen door $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$

3.3 De model-geïmpliceerde covariantie matrix

- eenmaal de parameters werden geschat kan men op basis van het paddiagram de bivariate covarianties/correlaties proberen te reconstrueren
- de zo bekomen covarianties/correlaties noemt men model-geïmpliceerde covarianties/correlaties
- in een gesatureerd model corresponderen deze exact met de geobserveerde covarianties/correlaties; in een niet-gesatureerd model is er vaak een (hopelijk zo kleine mogelijke) discrepantie tussen de voorspelde en geobserveerde covarianties/correlaties (= residuals)

4 Implementatie en interpretatie van padanalyse in lavaan

```
library(lavaan)

semmodel1<-'fitness~exercise+hardiness
            stress~hardiness+exercise+fitness
            illness~fitness+stress+exercise+hardiness'

fit1<-sem(semmodel1,data=data1)

summary(fit1)
```

lavaan-output

```
   Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         12

  Number of observations                           373

Model Test User Model:

  Test statistic                                 0.000
  Degrees of freedom                                 0

Parameter Estimates:

  Standard errors                            Standard
  Information                                Expected
```

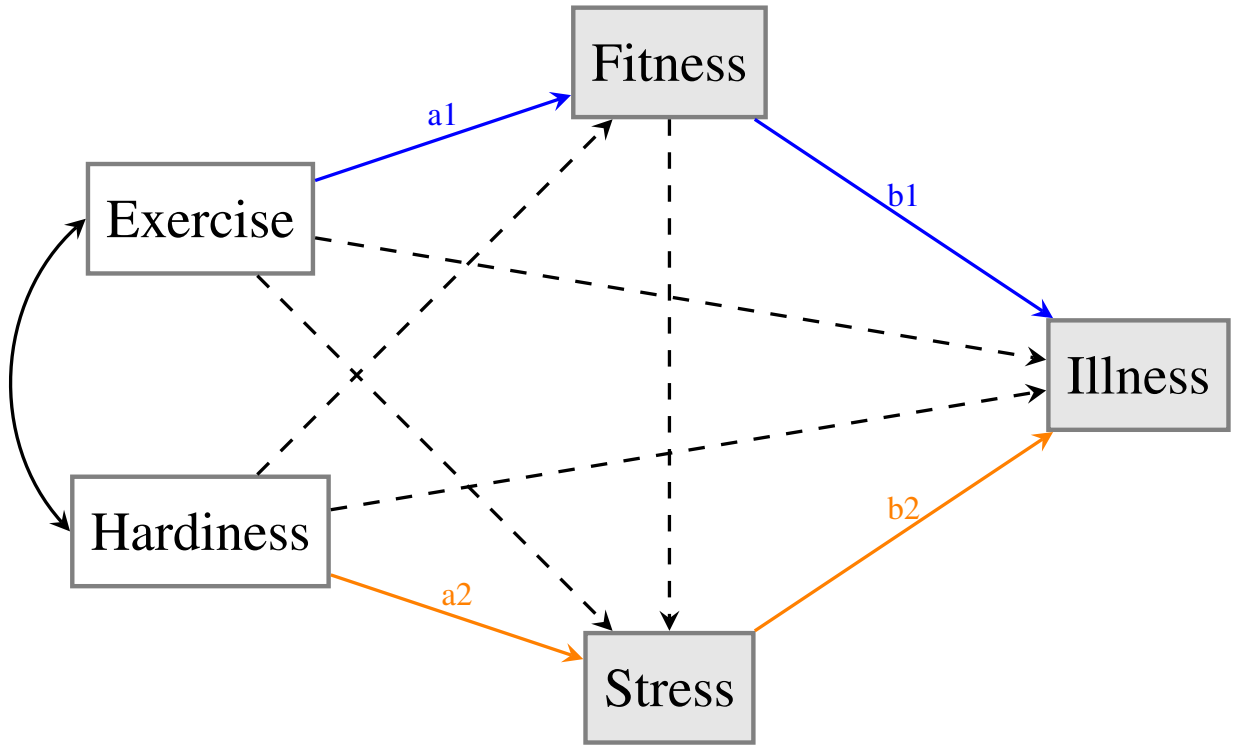
```
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  fitness ~
    exercise          0.217    0.026    8.260    0.000
    hardiness         0.079    0.046    1.721    0.085
  stress ~
    hardiness        -0.393    0.089   -4.433    0.000
    exercise         -0.014    0.055   -0.261    0.794
    fitness          -0.198    0.099   -1.995    0.046
  illness ~
    fitness          -0.442    0.087   -5.074    0.000
    stress            0.271    0.045    6.008    0.000
    exercise          0.032    0.048    0.664    0.507
    hardiness        -0.121    0.079   -1.532    0.126

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .fitness        1136.158   83.195   13.657    0.000
   .stress         4181.001  306.155   13.657    0.000
   .illness        3178.984  232.782   13.657    0.000
```

- gesatureerd model:
  - aantal datapunten: p=5, q=2: p(p+1)/2-q(q+1)/2=12
  - aantal parameters: 9 coëfficiënten en 3 varianties
- nagenoeg identieke schatters (en p-waarden) voor padcoëfficiënten als voor regressiecoëfficiënten in lineaire regressiemodellen
- varianties endogene variabelen $\approx$ gekwadrateerde residuele standard errors in lineaire regressiemodellen
- maar wat is het indirect effect van exercise/hardiness op illness?

Indirecte effect van fysieke activiteit en veerkracht op ziekte



Wat is het indirect effect van exercise op illness via fitness?
Wat is het indirect effect van hardiness op illness via stress?

```
semmodel2 <- 'fitness ~ a1*exercise + hardiness
              stress ~ a2*hardiness + exercise + fitness
              illness ~ b1*fitness + b2*stress + exercise + hardiness
              ie1exerc:=a1*b1
              ie2hard:=a2*b2'

fit2 <- sem(semmodel2, data = data1path)
summary(fit2)

...
Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)
    ie1exerc         -0.096    0.022   -4.323    0.000
    ie2hard          -0.107    0.030   -3.567    0.000
```

We vinden significante indirecte effecten:

- het indirect effect van exercise op illness via fitness: $-0.096(z = -4.323, p < .001)$
- het indirect effect van hardiness op illness via stress: $-0.107(z = -3.567, p < .001)$

Referenties

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling,* 6, 15.

Loeys, T., Moerkerke, B., & Vansteelandt, S. (2015). A cautionary note on the power of the test for the indirect effect in mediation analysis. *Frontiers in Psychology, Quantitative Psychology and Measurement.*

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software,* 48(2), 1-36.

Rosseel, Y. (2013). *Cursus Onderzoeksmethoden II.*

Roth, D.L., Wiebe, D.J., Fillingim, R.B., & Shay, K.A. (1989). Life events, fitness, hardiness, and health: A simultaneous analysis of proposed stress-resistance effects. *Journal of Personality and Social Psychology, 57*, 136–142.